University of Zurich[UZH]

Bachelor's thesis
for the degree of
Bachelor of Arts
at the Faculty of Arts and Social Sciences
of the University of Zurich

Measuring Reading Comprehension and Text Comprehensibility Using a Touchscreen Application

Author: Andreas Säuberli
Student ID: 17-705-161

Supervisor: Dr. Sarah Ebling
Department of Computational Linguistics

Date of submission: June 1, 2021

Abstract

Being able to measure reading comprehension is relevant for a variety of research questions and applications, for example testing a person's reading comprehension skills or evaluating the comprehensibility of texts in easy-to-read language. Often, readability formulas based on superficial linguistic features or perceived comprehensibility ratings are used, as more accurate and objective comprehension tests tend to be expensive, time-consuming and sometimes difficult to implement for subjects with reading difficulties. In this thesis, I propose using computer-based testing with touchscreen devices as a means to simplify and accelerate data collection using comprehension tests, and to facilitate experiments with less proficient readers. I demonstrate this by designing and implementing a mobile touchscreen application and validating its effectiveness in an experiment with a small sample of people with intellectual disabilities. The results suggest that there is no difference between measuring comprehension using the application and traditional paper-and-pencil tests. In the future, such an application could be used to collect more reliable and authentic data about information accessibility in texts with diverse target groups, but more extensive testing and usability improvements are still necessary.

Zusammenfassung

Leseverständnis messbar zu machen ist relevant für eine Vielzahl von Forschungsfragen und Anwendungen, beispielsweise für die Prüfung der Leseverständnisfähigkeiten einer Person oder die Evaluierung der Verständlichkeit von Texten in leichter Sprache. Oftmals werden Lesbarkeitsformeln basierend auf oberflächlichen sprachlichen Merkmalen oder subjektive Verständlichkeitsbewertungen verwendet, da genauere und objektivere Verständnistests tendenziell teuer, aufwendig und manchmal schwer implementierbar für Personen mit Leseschwierigkeiten sind. In dieser Arbeit schlage ich computerbasierte Tests auf Touchscreen-Geräten vor als Mittel, die Datenerhebung mit Verständnistests zu vereinfachen und zu beschleunigen, und Experimente mit Teilnehmern mit geringerer Lesefertigkeit zu ermöglichen. Um dies zu demonstrieren, entwerfe und implementiere ich eine mobile Touchscreen-Applikation und überprüfe deren Effektivität in einem Experiment mit einer kleinen Stichprobe von Personen mit kognitiven Beeinträchtigungen. Die Resultate deuten darauf hin, dass es keinen Unterschied gibt zwischen Verständnismessungen mithilfe der Applikation und traditionellen Tests auf Papier. In Zukunft könnte eine solche Applikation dazu verwendet werden, zuverlässigere und authentischere Daten über die Zugänglichkeit von Informationen in Texten mit diversen Zielgruppen zu erheben, wobei aber noch umfangreichere Tests und Verbesserungen der Benutzerfreundlichkeit notwendig sind.

Acknowledgments

First and foremost, I would like to thank Sarah Ebling, who motivated me to take up this challenge, supported me during the entire project and was the best supervisor I could have wished for.

I would also like to thank Silvia Hansen-Schirra and Silke Gutermuth for sharing their expertise in easy-to-read language reception and for their guidance on the experiment design, and Laura Schiffl and Silvana Deilen for their valuable feedback on the task implementations and the project in general.

I am also indebted to Franz Holzknecht for helping me with the experiment design and data analysis.

I am grateful to collaboration partners at *capito*/*atempo*, especially Ursula Semlitsch for her experience with the participant group and her diligent work in recruiting participants and conducting the experiment. I also owe special thanks to all the study participants.

Last but not least, I would like to thank my fellow students at the University of Zurich and my family for motivating and enduring me throughout the highs and lows of the project.

Contents

List of Figures

List of Tables

List of Acronyms

APA Austria Press Agency 13

API application programming interface 11

CEFR Common European Framework of Reference for Languages 6

EEG electroencephalography 5

GPL GNU General Public License 11

LL Leicht Lesen 6, 13

MVP minimum viable product 1, 11

PWA progressive web application 9, 11

QR Quick Response 9

UI user interface 9, 11, 22

1 Introduction

Quantifying how well a text is understood by a reader is difficult. Nevertheless, many research questions and applications heavily depend on reliable measurements of comprehension, for example when evaluating and estimating the quality of automatic text simplification systems, verifying information accessibility for people with reading difficulties, finding appropriate language material for use in education, and for numerous research areas in psycholinguistics. Besides the methods of measurement being difficult to implement, large sample sizes are usually needed due to high between-subject variability, and data collection is slow and expensive. This is particularly true for research questions involving populations with lower reading proficiencies, such as the typical target groups of easy-to-read language. At the same time, many members of these target groups are able to use devices like personal computers or smartphones effectively and independently. This raises the question: Is there a way to leverage accessible technology to accelerate and simplify the measurement of reading comprehension with a wide range of representative readers?

This thesis takes a closer look at these issues and proposes a way to improve experimental data collection using touchscreen-based testing. It also presents initial evidence to assess the effectiveness of this method.

The thesis is divided into three parts. The first part (Chapter 2) is an overview of existing commonly used approaches to measuring reading comprehension and text comprehensibility, and highlights their advantages and disadvantages. In the second part (Chapter 3), I present the design and implementation of a minimum viable product (MVP) for a touchscreen application with the goal of facilitating comprehension measurement with larger sample sizes and target groups with more accessibility needs. The last part (Chapter 4) is an experimental validation of the application with participants with intellectual disabilities, comparing it with traditional paper-and-pencil methods.

2 Reading comprehension and text comprehensibility

Reading comprehension is a highly complex psychological process which involves many different cognitive skills. Additionally, comprehension cannot be directly and objectively observed, but has to be inferred from reader responses or behavior assumed to be affected by it. Consequently, the results of reading comprehension assessments depend not only on the skills of the reader and the nature of the text, but also on the method of measurement (Fletcher, 2006).

After some terminological clarifications, this chapter provides a (non-comprehensive) overview of the common methods used to measure reading comprehension and text comprehensibility (Section 2.2) and a brief description of easy-to-read language as a use case (Section 2.3).

2.1 Terminology

As these topics have been actively researched for a very long time and in many different fields from education and psychology to computational linguistics, the terminology related to reading comprehension and text comprehensibility is quite diverse and sometimes inconsistent. The following list is an attempt at disambiguating some of the most common terms and explains their usage in this thesis.

(Reading) comprehension refers to an individual's ability to (correctly) understand a text, or the corresponding cognitive processes. Measuring comprehension means finding out how well an individual is able to understand a text (usually independently of the text's complexity).

Comprehensibility is a property of a text and denotes how well a text can be understood (usually by a particular target group or the average reader), i.e. the effect of the text's complexity or difficulty on the reader's comprehension.

Readability is mostly used synonymously to comprehensibility, but strongly influenced by the phenomenon of readability formulas (Bailin and Grafstein, 2016, p. 1). It may also involve typographical legibility or how interesting the content is to the reader (Charzyńska and Dębowski, 2015, p. 125). I will avoid this term except in the context of readability formulas.

Text difficulty is sometimes used as an antonym (and **text accessibility** as a synonym) for comprehensibility or readability (Fulcher, 1997).

2.2 Approaches to measuring comprehension and comprehensibility

This section outlines the most common methods applied in research for measuring reading comprehension and/or text comprehensibility, grouped into three main categories, and presents examples in previous literature.

2.2.1 Readability algorithms

Readability algorithms are a way of computationally estimating the comprehensibility of a text. The most well-known algorithms are the traditional readability formulas developed from the 1950s to the 1970s, which are still sometimes used in education and research, mainly due to their simplicity. These use superficial linguistic features such as sentence, word and syllable counts (e.g. Flesch Reading Ease, Flesch, 1979; the Gunning FOG formula, Gunning, 1952; or the LIX formula, Björnsson, 1968) and sometimes pre-defined lists of "easy" words (e.g. the Dale-Chall formula, Dale and Chall, 1948) to predict comprehensibility on a one-dimensional scale, often mapped to specific academic grade levels. Recently, more sophisticated algorithms have been developed, including statistical approaches leveraging language models (e.g. Collins-Thompson and Callan, 2004) and more complex syntactic, semantic and discourse coherence related features (e.g. Graesser et al., 2004; Pitler and Nenkova, 2008). Feng (2010, p. 23–32) provides a more comprehensive overview of these types of approaches.

The sheer number and the vast variety of algorithms developed within the past 70 years is already telling that these are not generic or universally applicable methods but have very limited use. For as long as they have existed, they have been criticized for oversimplifying the complex topic of reading comprehension, ignoring reader-specific features, and a lack of empirical basis (Anderson and Davison, 1986; Fulcher, 1997; Bailin and Grafstein, 2001). Due to this oversimplification, they cannot be used to guide a writer in producing comprehensible texts, but only as rough estimates which tend to correlate with certain aspects of comprehensibility (Klare et al., 1963).

Because of these limitations, Bailin and Grafstein (2016) highlight the importance of empirical evidence to justify and validate readability algorithms. The summary by Benjamin (2012, p. 81–82) shows that for many approaches, there is no or insufficient evidence of validity, and that the used validation methods differ substantially. In many cases, readability algorithms are evaluated purely on their ability to predict grade levels or language levels based on expert judgments, rather than evidence from humans within the target group they are meant to model. When evidence involving humans is available, it is usually through comparison with one or more of the approaches described in Sections 2.2.2 and 2.2.3.

2.2.2 Perceived comprehensibility ratings

A more human-centered approach is to ask a sample of people to read a text and judge its comprehensibility, most often on a single linear scale, for example using Likert ratings. Kandula and Zeng-Treitler (2008) use ratings both for creating and validating a gold standard for measuring comprehensibility of health texts and argue that this method is more robust than readability formulas and requires a smaller number of subjects than cloze tests. Pitler and Nenkova (2008) attempt to capture multiple aspects of comprehensibility by collecting ratings for four different questions about each text, but the responses turn out to be all mostly the same. They also show that the surface metrics used in traditional readability formulas are not particularly good predictors of these ratings. Alonzo et al. (2021) suggests that subjective ratings might be more effective at discriminating between simple and complex texts than objective metrics from comprehension tests, at least in the case for deaf and hard-of-hearing readers.

Using comprehensibility ratings seems particularly popular in evaluating automatic text simplification, where most of the time, experts – or at least highly proficient raters – are used instead of readers from the text's actual target group (e.g. Xu et al., 2016; Sulem et al., 2018). Reasons for this may be the assumption that they are capable of making more objective judgments, or that they are more readily available in academic environments. Apart from the danger that these results may not be transferable to the real target population of the text, expert ratings have been shown to exhibit low inter-rater reliability due to different experts applying very different criteria (Fulcher, 1997).

2.2.3 Comprehension tests

This category encompasses all tests which measure reading behavior or involve tasks with the goal of inferring "objective" (as opposed to perceived) comprehension, sometimes referred to as "actual readability". Since comprehension cannot be observed directly, these tests have to make some psychological assumptions about the readers, for example *The better the reader understands a text, the more comprehension questions they can answer correctly*, or *The better the reader understands a text, the faster they can read the text.* Conducting these tests and interpreting their results is more difficult, because in reality these assumptions are not as straightforward, and many other effects can influence the measurements.

Some of the most commonly used tasks are multiple-choice or true/false comprehension questions (e.g. Huenerfauth et al., 2009; Leroy et al., 2013; Charzyńska and Dębowski, 2015; Vajjala and Lucic, 2019; Alonzo et al., 2021) and open-ended or free recall questions (Leroy et al., 2013; Charzyńska and Dębowski, 2015; Vajjala et al., 2016). As Huenerfauth et al. (2009, p. 8–9) points out, factors such as the participant's interest, prior knowledge, and understanding of the questions can cause a lot of noise in the measurements for these tasks.

Other popular tasks include cloze tests (e.g. Charzyńska and Dębowski, 2015; Redmiles et al., 2019) and self-paced reading (e.g. Crossley et al., 2014; Fulmer et al., 2015). More

rarely, and especially when specific linguistic features are investigated in isolation, data from lexical decision tasks, eye-tracking or electroencephalography (EEG) are also used (e.g. Pappert and Bock, 2019; Vajjala et al., 2016; González-Garduño and Søgaard, 2017; Gutermuth, 2020; Yuan et al., 2014).

A clear disadvantage is that all of these require large samples of participants, and most of them have to be designed by experts. Some measurement methods are also somewhat obtrusive or force participants to deviate from their natural reading behavior. Yet they are thought to be more representative of true comprehension, and a combination of these tasks are often used to validate simpler and less expensive measurement approaches such as expert ratings or readability formulas (e.g. Vajjala et al., 2016; Charzyńska and Dębowski, 2015).

2.2.4 Summary and discussion

We have seen that all of the approaches above have advantages and disadvantages, and all of them can be useful in different situations. Table 2.1 gives a summary of some of the main criteria distinguishing the three categories presented. It is also important to note that there is no standardized way of validating any of these methods, and different sources use different comparisons to argue for their approach, since there is no way of measuring true comprehension.

Finding methods to measure comprehension can be thought of as looking for windows into the reader's mind: The windows are always a bit dirty and add noise to the observed signal, and a single window can only illuminate a small fraction of the complex cognitive processes involved. The key is to find the right windows for a given research question and using them in conjunction to get a more complete picture.

Criterion	Readability algorithms	Subjective ratings	Comprehension tests
Is it easy to apply?	+	+	−
Is it easy to interpret?	+	+	−
Is it inexpensive?	+	−	−
Is it objective?	−	−	+
Is it unobtrusive?[1]	+	±	−
Can it capture between-reader differences?	−	+	+

Table 2.1: Simplified overview of selected (dis)advantages of different approaches for measuring reading comprehension and text comprehensibility

[1] Here, "unobtrusive" refers to the interference with participants' natural reading behavior, for example due to gaps inserted in the text or eye-tracking equipment.

2.3 Easy-to-read language

To highlight how measuring comprehension is relevant to real-life applications, and to give some context for the experiment described in Chapter 4, this section will briefly outline what easy-to-read language is and how it relates to measuring comprehension.

Easy-to-read language is a constructed variety of standard language with the aim of making information more accessible, particularly, but not exclusively, for target groups such as people with intellectual disabilities, prelingually deaf people or non-native speakers. It usually features reduced complexity on lexical and syntactic levels, additional explanations for difficult concepts and clearly structured layout to increase comprehensibility for these target groups.

Such varieties have been proposed in several languages, with guidelines about which linguistic and graphical elements to use or avoid. For German, there are multiple well-documented easy-to-read varieties, including *Leichte Sprache*, which is being developed by a dedicated association and officially indorsed by the German government, and *Leicht Lesen (LL)*, which is an initiative by *capito*/*atempo*, the largest provider of human text simplification services for German. The latter itself comprises three different levels of simplification (A1, A2 and B1[2]).

Currently, there is no single widely accepted standard, and research about these varieties is still lacking. Clearly, it is crucial that easy-to-read language be evaluated with appropriate methods and relevant target groups, especially if its usage is to be recommended or enforced by legislative authorities. This is also one of the motivations behind the development of the app presented in Chapter 3.

[2]Although these labels are borrowed from the Common European Framework of Reference for Languages (CEFR), they are not entirely comparable to foreign language proficiency, as they focus on different target groups.

3 *Okra*: a touchscreen application for comprehension testing

This chapter describes the goals, design decisions and technical implementation of the *Okra* application, which implements some of the approaches mentioned in Chapter 2. I will start with some more general remarks about the advantages and disadvantages of computer-based testing to put the decisions into perspective.

3.1 Aspects of computer-based testing

Use of computer-based applications with the purpose of testing specific human abilities has been on the rise in the past decades. The advent of widespread use of personal mobile devices and the consequently increased technical familiarity of the general public fosters this development even more. Computer-based testing has been successfully studied and applied not only in education, but also in identifying neurological disorders (e.g. Boukhvalova et al., 2018; Jongstra et al., 2017), cognitive assessment (e.g. Timmers et al., 2014), psychological and psycholinguistic research (e.g. Stoet, 2017; Valliappan et al., 2020; Dufau et al., 2011) and others.

The most obvious advantage of computer-based (and especially smartphone-based) testing for research purposes compared to paper-and-pencil methods is the possibility to gather more data more quickly, which is an essential issue in any experiment which relies on human responses. Not only can it reduce administration and equipment costs, but it makes it possible to move the experiment out of the laboratory if the research question allows for it. Even though less controlled conditions may be detrimental to data quality, this effect may be counterbalanced by an increased reach of participants (e.g. through snowball sampling), and it may even be desirable to collect data in a more natural day-to-day environment (Timmers et al., 2014, p. 137).

Moreover, personal computers and mobile devices offer some benefits which cannot easily be achieved otherwise – with or without laboratory conditions. The sense of familiarity associated with one's personal device lowers the threshold for users to participate and increases motivation, not to mention the wide range of available options for gamification. As people are less emotionally invested in computers than humans, they are much more open towards receiving computer-mediated feedback, and the user's self-esteem has a significantly smaller effect on performance than when feedback is given by a human (Kluger and Adler, 1993). In addition, modern smartphone and tablet devices offer many built-in personalization settings and accessibility features (e.g. font size customization or color correction), which can be used to create familiar and accessible

user experiences for diverse target groups without losing control over stimulus presentation. Ubiquitous access to the internet allows remotely conducted experiments in a more dynamic way, for example by distributing and adapting experimental items on-the-fly, reminding participants with push notifications (especially in the case of longitudinal studies) and monitoring responses in real time.

But there are also significant disadvantages, which must not be overlooked. Although smartphones and tablets feature an exceptionally diverse set of input and output modalities (sensors, cameras, displays, speakers etc.), these are not always comparable across devices due to hardware or software differences (e.g. different screen sizes, operating systems or sensor precisions). For methods such as eye-tracking for reading, which requires very high image resolutions, the hardware found in affordable devices is still insufficient (Valliappan et al., 2020). Finally, conducting experiments outside of laboratory conditions using participants' personal devices lead to more external factors such as distractions, and possibly a higher cancellation rate.

These aspects should guide application designs for computer-based testing and must be kept in mind when planning any computer-mediated experiment.

3.2 Goals and design requirements

The goal is to create an application which can be used by participants to complete experimental tasks related to reading comprehension on a touchscreen device. It should make use of the advantages and mitigate the disadvantages mentioned in the previous section in the best way possible, and find an appropriate balance between user experience and usefulness of the collected data. More specifically, I define the following requirements to guide the design and implementation, for the three main groups of people who will interact with the application.

Participants should be able to:

(P1) use the app comfortably, even with mild to moderate intellectual and/or physical disabilities

(P2) understand what they are asked to do

(P3) complete tasks without being overwhelmed by information or actions

(P4) take breaks or cancel a task whenever they need or wish to

(P5) find motivation and encouragement to start and complete tasks

(P6) use the app both on their own devices and on unfamiliar devices in laboratory conditions

(P7) be assured that their personal data is safe

Researchers should be able to:

(R1) configure tasks to adapt them to their experiment design and participant group

(R2) conduct experiments with any type of procedure

(R3) reliably get data which is accurate, precise and comparable across participants

(R4) get response data immediately and interfere with the procedure if necessary

Developers should be able to:

(D1) easily customize the back-end implementation to adapt it to any infrastructure and data security requirements

(D2) automate data analysis and visualization

(D3) easily extend the application to add different types of tasks

(D4) easily maintain the application across different platforms and operating systems

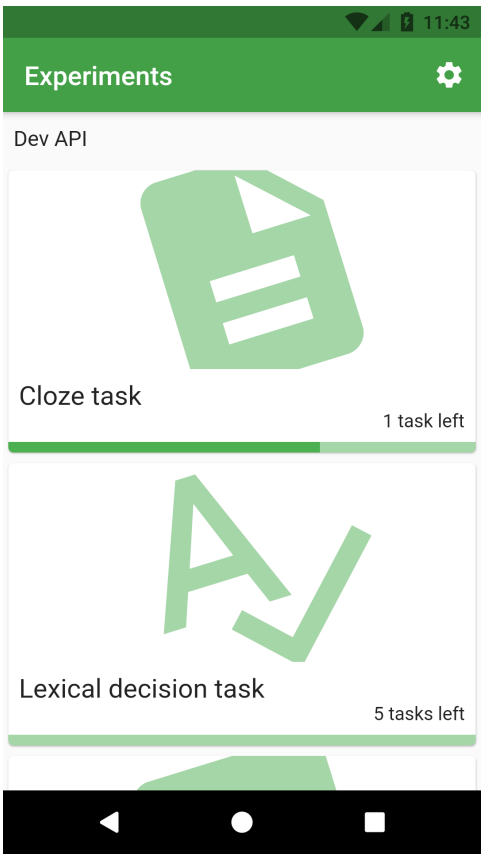
3.3 Implementation

This section describes the application's user interface (UI) and its underlying architecture, referencing the corresponding requirements listed above whenever applicable.

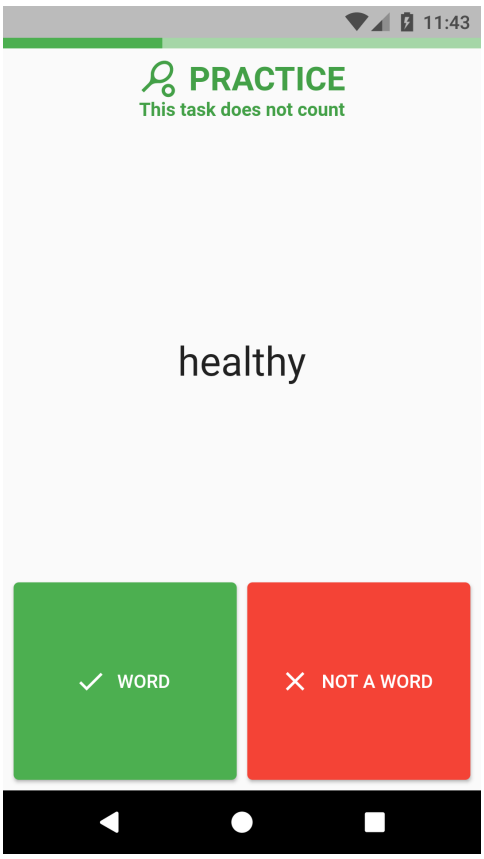
Okra is designed to be used by participants as independently as possible (P6): It is a cross-platform application which could be made available on mainstream app stores or as a progressive web application (PWA) for easy installation. Participants need to register using credentials given to them by a researcher in order to receive experiments to participate in. This can be done by participants themselves by scanning a Quick Response (QR) code, or by a researcher if the experiment is conducted in person. The application does not ask the participant for any personal information and does not store any data except the anonymous credentials entered to register (P7).

Participants can then see the tasks currently available to them, view their instructions and start completing them. If the experiment design includes a practice task, participants are required to complete it first and are allowed to repeat it if needed (P2). While working on a task, additional UI elements are hidden to draw full focus to the relevant stimuli and input elements. Tasks are also implemented in a way that only the minimum amount of information is shown at the same time (P3). A subtle progress bar is shown to indicate how long the task will take to finish (P5). After finishing a task, an encouraging message appears and participants are given the option to continue immediately with the next task or take a break (P4). Aborting a task prematurely is possible, but discouraged by a dialog box. Some example screenshots can be seen in Fig. 3.1.

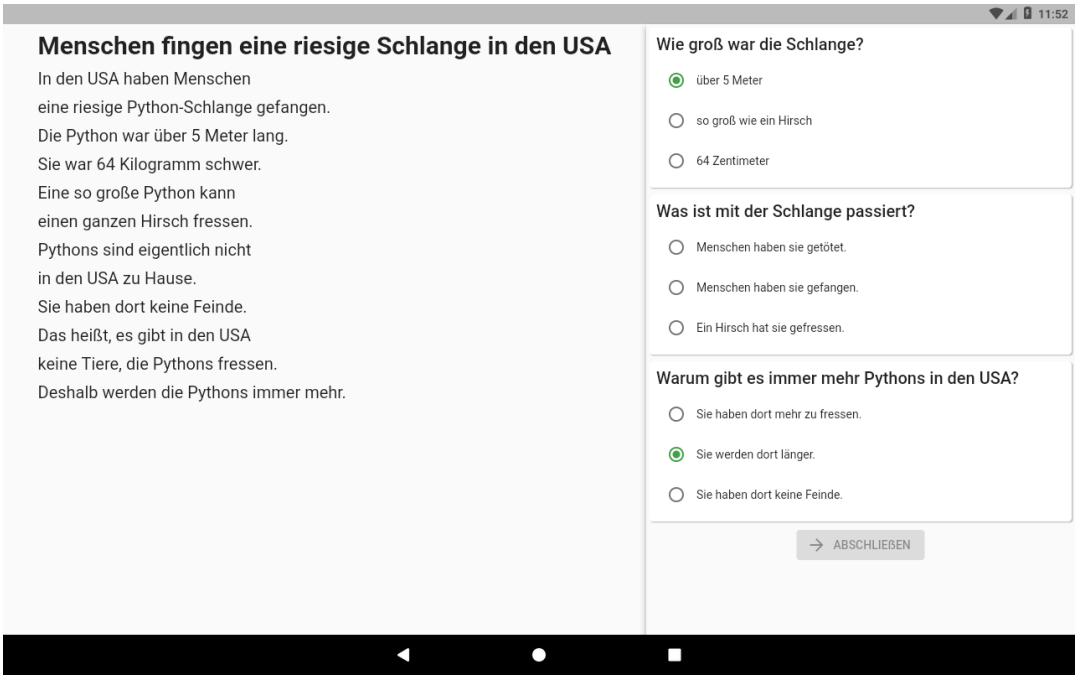
Okra as a client-side application receives experiment configurations and stimulus data from a server, presents the task (e.g. a multiple choice or lexical decision task) to the participant, records the participant's response data (e.g. answers or reaction times) and



(a) List of available experiments on an Android phone



(b) Practice trial of a lexical decision task on an Android phone



(c) Comprehension question task with a German text on an Android tablet

Figure 3.1: Screenshots of *Okra* with example tasks

sends them back to the server. Client and server communicate using a web application programming interface (API). While the task types and UI are implemented in the client, task distribution and data storage is entirely up to the server implementation. This strict separation of client and server responsibilities allows anyone conducting experiments to use a custom server implementation to meet individual privacy requirements (D1), monitor and analyze results immediately (R4) and distribute tasks in responsive ways, for example to automatically adapt task difficulty to a participant's performance in previous tasks (R2).

The application is written using *Flutter*[1], a modern cross-platform UI development kit, which allows developing native Android, iOS or desktop applications and PWAs from a single codebase and provides fast and consistent graphics rendering across devices and platforms (D4). Additionally, it facilitates UI design according to material design guidelines[2], provides extensive support for accessibility features (P1), and adapts the visual appearance to the operating system's native standards to make the application feel familiar to users, without losing pixel-level control over stimulus presentation and timing (R3).

The client-side source code and documentation is released[3] under GNU General Public License (GPL) 3.0. The web API is defined and documented using the widely used *OpenAPI*[4] specification standard to facilitate custom back-end implementations (D1). A simple example or reference implementation in Python using the web framework *Django*[5] is also available[6].

3.4 Tasks

Each task type is implemented in a modular way and with its own configuration settings to make it easy to create new and extend existing task implementations (D3). The MVP version of the application used in the validation experiment includes some tasks testing comprehension and some testing more low-level cognitive and/or motor skills:

- A multiple-choice cloze test, where a short segment of text with a single gap is shown at a time.
- A lexical decision task, where a single string of characters is to be judged as a word or a non-word to test word recognition.[7]
- A reading task with multiple-choice questions.[7]
- An n-back task testing working memory, first introduced by Kirchner (1958).

[1] `https://flutter.dev/`
[2] `https://material.io/`
[3] `https://github.com/saeub/okra`
[4] `https://swagger.io/specification/`
[5] `https://www.djangoproject.com/`
[6] `https://github.com/saeub/okra-server-example`

- A word-picture-matching test, where participants choose the matching picture for the displayed word, as described by Deilen (2020).
- A reaction time test, where images appear on screen and participants tap them as quickly as possible.[7]
- A simple clone of the electronic memory skill game *Simon*, where participants remember an increasingly long sequence of buttons to press.[7]

For the most part, variables affecting difficulty can be configured for each task, to match the level of participants. Whenever applicable, visible feedback about answer correctness can also be configured (R1). For each task type, events such as screen tapping and timed stimulus appearances are recorded with timestamps and sent to the server after completion, such that all relevant user interactions can be precisely retraced (R3).

[7]These task types were used in the validation experiment and are described in more detail in Section 4.3

4 Experimental validation and exploration

To verify the effectiveness of the application in measuring text comprehension, we conducted a small-scale experiment with people with mild to moderate intellectual disabilities in collaboration with a partner at *capito*/*atempo*. We compared comprehension measurements collected using the touchscreen application with a more traditional paper-and-pencil test to find out whether there are differences depending on modality.

Another goal is to get a first impression of the applications's usability and accessibility, and to explore further (dis)advantages and aspects to take into account when considering this modality over non-digitized testing. To this end, and to get more insight into participant's cognitive and motor skills, the study also includes three other tasks, all of which are completed in the application.

4.1 Participants

The 16 participants were part of an educational program for people with learning difficulties and disabilities at *atempo* in Austria and agreed to take part in the study on a voluntary basis. They were deemed by their program instructor to be part of the target group for simplified texts around the level given in the presented material (A2). All participants were informed about the study and signed a consent form written in easy-to-read German ahead of time, which they were all legally eligible to sign themselves. They were compensated with cash for participating in the study.

There were eight male and eight female participants, their mean age was 26.1 (youngest: 18, oldest: 38, median: 26) and all were native German speakers, one being bilingual. Based on self-reports, eleven were visually impaired (corrected with glasses), none were colorblind. Most of them were very familiar with touchscreen devices, 14 of them reporting daily smartphone use, two of them weekly. Self-reported reading behavior (e.g. reading texts in newspapers, books or on the internet) differed considerably, four participants reading every day, eight once per week or more, five less than once per week.

4.2 Materials

The texts presented to the participants are taken from the Austria Press Agency (APA) corpus described in Säuberli et al. (2020) consisting of short Austrian news articles translated into simplified German of level A2 and B1 following LL guidelines (cf. Section 2.3).

Text	#sentences	#words	LIX[1]
(practice)	3	30	51.0
unemployment	9	90	43.3
school-vacation	8	79	35.5
fish	15	122	23.0
glyphosate	9	86	40.9
murder	9	88	32.5
snake	9	70	29.4
ski-championship	8	63	48.0
film-festival	8	70	28.7

Table 4.1: Texts used in the experiment

I manually picked eight texts (plus one shorter text for practice trials) at level A2 such that there is a variety of topics and enough opportunities for developing comprehension questions. All texts were published in 2018 or 2019 and do not cover currently very relevant or generally known issues or events, to reduce the effect of prior knowledge. All images were removed from the texts as they are not necessary for comprehension, but the original layout with short line lengths was preserved. Table 4.1 shows some statistics about the selected texts, including the LIX readability index (Björnsson, 1968).

I designed three comprehension questions per text with three answer choices each. This type of multiple-choice question has a lower chance of lucky guesses than yes/no questions and avoids involving other skills like writing or speaking. The questions are designed according to these rules:

- No answer choices can be excluded by merely applying common knowledge or logic without having read the text.
- The three questions are independent of each other, i.e. the answer of one of the questions does not imply the answer to one of the other questions.
- The three answer choices are mutually exclusive.
- Incorrect answer choices use "plausible" distractors, e.g. numbers or words that appear or could contextually appear in the text.

I also continuously integrated feedback from experts familiar with the target group and an expert in language testing to make sure the questions themselves are easy to understand, appropriately difficult to answer and indicative of true comprehension. The resulting questions are mainly literal and re-organizational in nature and do not require inferring implicit information, since the latter question type is significantly more difficult to answer (Vajjala and Lucic, 2019; Fajardo et al., 2014).

[1]Calculated using https://www.psychometrica.de/lix.html. A value below 40 is common for literature for children, 40–50 for fiction, 50–60 for nonfiction, above 60 for technical literature

4.3 Method

After obtaining approval for the procedure from the ethics committee and after a pilot study with two participants, the experiment was administered by an employee at *atempo* over the course of nine days in May 2021. Each participant attended two sessions on separate days, most sessions taking between 15 and 30 minutes, to avoid fatigue and to capture within-subject variance more accurately.

At the beginning of the first session, the participant answered some questions about their personal data, reading behavior and use of technology, which were asked orally by the experimenter. Then, the participant was given four texts to read, two on paper and two using *Okra* on a tablet (Apple iPad 2018, 9.7 inches). After reading a text, the participant rated the difficulty of the text on a five-point horizontal Likert scale of colored smiley faces, with extremes labeled as "very difficult" and "very easy" ("sehr schwierig" and "sehr einfach" in German; positive sentiment on the right[2]). After this, they received the comprehension questions and were instructed to answer them, allowing them to look at the text again if they wanted to. Finally, they rated the difficulty of the questions in the same way and also answered the question "How much did you enjoy this task?" ("Wie viel Spaß hat Ihnen diese Aufgabe gemacht?" in German) on the same type of Likert scale, with extremes labeled as "not at all" and "very much" ("gar keinen Spaß" and "sehr viel Spaß" in German), before moving on to the next text. The time taken to read the text and to answer the questions was measured, either by the application directly or by the experimenter for texts read on paper. Before the participant read the first text, and whenever the condition (paper/app) was switched, they were presented again with written instructions in easy-to-read German, and were asked to complete a practice task including reading a shorter text, two comprehension questions and the usual ratings, to re-familiarize them with the procedure and the material. Typography and layout of the texts were matched across the two conditions as well as possible. After having read two texts (not including the practice text), the experimenter offered a short break and a glass of water.

At the beginning of the second session, the participant was asked to solve three low-level cognitive tasks in *Okra* on the tablet, in this order:

1. A reaction time test, where a red balloon appeared at a random position on the screen and the participant was instructed to tap the balloon as quickly as possible. After they did, the next balloon appeared after a random delay between 0.0 and 1.0 seconds. In total, 10 balloons were presented and the reaction times measured.

2. A lexical decision task, where a single string of characters was presented at a time and the participant was instructed to press the button labeled "WORD" as quickly as possible if the word is a real German word or otherwise press the button labeled "NOT A WORD". I used the *DeReWo* word frequency list (Perkuhn et al., 2009)

[2]Maeda (2015) finds horizontal computer-based rating scales to be faster to complete but with a left-side selection bias. Franzen (2014, p. 672) argues that the order of rating scales does not play a crucial role. The decision to use a horizontal scale here was mainly based on usability considerations.

to select words between three and ten characters long within the top 5000 words and the *Wuggy* software (Keuleers and Brysbaert, 2010) to generate pseudowords. Ten words and ten pseudowords were presented in randomized order.

3. A short-term memory game, where four differently colored buttons (each with a different geometric shape on it to aid memorization) were visible in a grid. The participant was instructed to memorize the sequence in which they lit up and repeat the sequence afterwards by pressing the buttons in the same order. The sequence started at length one with an element being added and the new sequence presented every time it was correctly repeated. The trial ended as soon as a button is incorrectly pressed.

Before each task, the participant read in-app instructions in easy-to-read German and completed a short practice task. The experimenter asked them to repeat the practice task in case they did not understand it. Participants were instructed to use their dominant hand for tapping. After completing all three tasks, four texts with comprehension questions and ratings were again presented in the same manner as in the first session. The order of the texts and conditions were randomized such that each participant read all eight texts and every text was read 8 times in each condition.

The same experimenter was present in the room in all sessions, sitting at a different table not directly in front of the participant, but still within their field of view. For tasks done on the tablet, the experimenter used a screen mirroring software to be able to see the tablet screen on a laptop and intervene in case of technical problems or when the participant did not understand a task.

4.4 Results

4.4.1 Comprehension questions and ratings

The response data from one of the participants reading the text `film-festival` on the tablet was lost due to a software bug, which was then immediately fixed, leaving data from 127 text readings and all the data from the three low-level cognitive tasks.

On average, participants had an accuracy of 2.19 out of three correctly answered questions per text (standard deviation: 0.41, lowest: 1.63, highest: 2.88). The mean accuracies between the two conditions did not differ significantly (app: 2.14, paper: 2.23 out of three). However, since the participants' language proficiencies as well as the difficulty of the questions may vary substantially, these raw accuracies are not necessarily fair or comparable. Modeling the data using Many-Facet Rasch Measurement (Linacre, 1989) can provide more insight into these aspects by mapping participant proficiency and item difficulty onto a common logit scale, while also taking into account the effects of other variables such as, in this case, whether the test was conducted on paper or on a touchscreen device. Figure 4.1 shows a Wright map of participant and item measurements fitted using the *FACETS* software (Linacre, 2020), where participants are ordered according to their proficiency (higher measure means more proficient) and items

according to their difficulty (higher measure means more difficult). A participant has a modeled probability of 0.5 of correctly answering an item with the same logit measure as theirs. If the item's measure is higher than the participant's, this probability is smaller than 0.5.

Question 2 of text `ski-championship` was answered correctly by all participants and is therefore uninformative with respect to proficiency. For the remaining data points, mean-square infit statistics are mostly close to 1.0, with participant infits ranging from 0.70 to 1.44 and item infits ranging from 0.75 to 1.38, which indicates that there are few unpredictable responses and supports the validity of the tests. However, as Fig. 4.1 shows, the questions tended to be too easy for many of the participants. The conditions have a separation and reliability of 0.00, which suggests that there is no difference in difficulty due to testing modality. A bias analysis also shows that there is no significant interaction between any of the items and the conditions (all $p > 0.17$).

Since each question has its own degree of difficulty, it is impossible to infer anything about the texts' comprehensibility from the multiple-choice responses alone. We can, however, compare the subjective Likert scale ratings to find out about perceived comprehensibility. To this end, I fit another Many-Facet Rasch Measurement model with a rating scale for each of the three ratings. These models take into account the facts that some participants may be more lenient than others and that they might not interpret the Likert scales as linear. The Wright maps in Fig. 4.2 show subjective difficulty/enjoyment (higher measure means easier/more enjoyable) and participant leniency (higher measure means more lenient). Many participants had a strong tendency towards very high ratings, especially for the text difficulty and enjoyment ratings, and some participants rated all texts with the same scores (four in the case of text and question ratings, and seven in the case of enjoyment ratings), which raises question about the usefulness of this data. All of the ratings correlate significantly with each other (Pearson's $r > 0.55$, $p < 0.001$), and none of them correlate with the comprehension question accuracy. There is, again, no separation between the two conditions and no significant interaction between texts and conditions.

The time taken to read the texts differs between the two conditions, as Fig. 4.3 shows. This duration was measured from when the participant first saw the text until they finished the text difficulty rating. Applying a linear mixed model with texts and participants as random effects shows that mean reading times on the tablet are significantly shorter than on paper (difference: 9.97 seconds, $p < 0.001$).

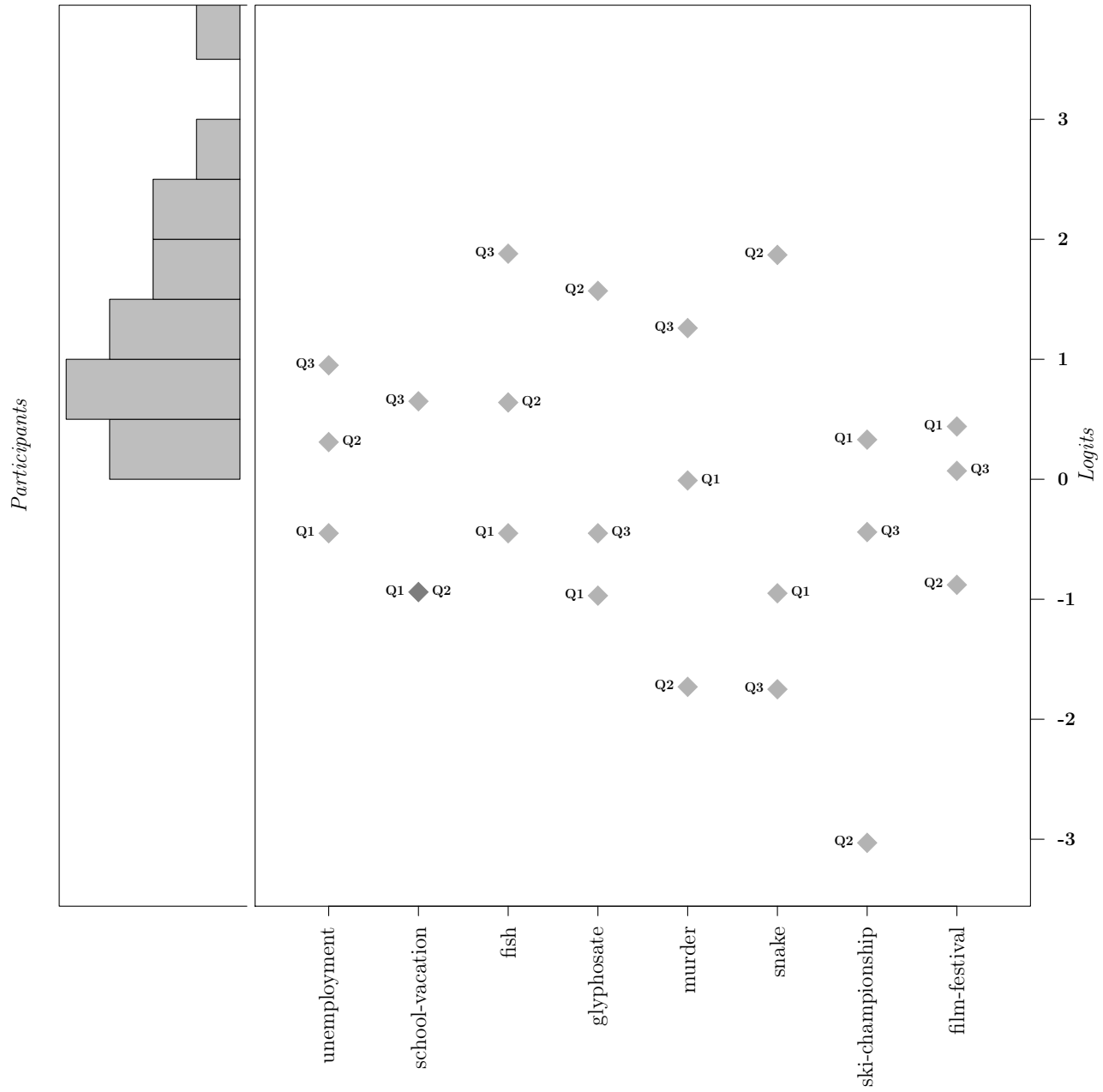
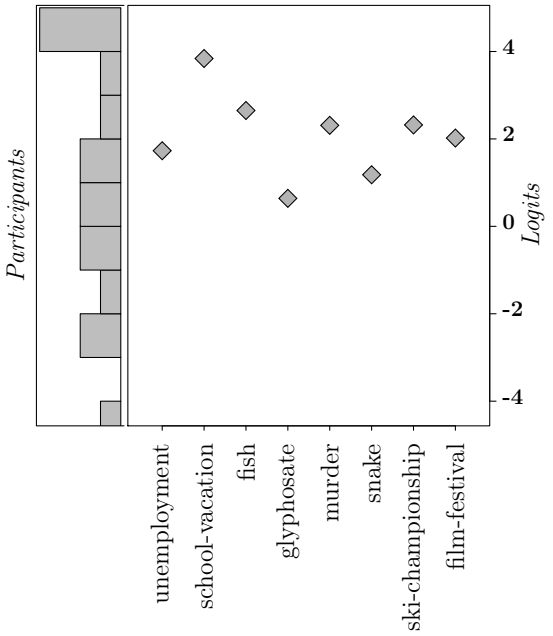
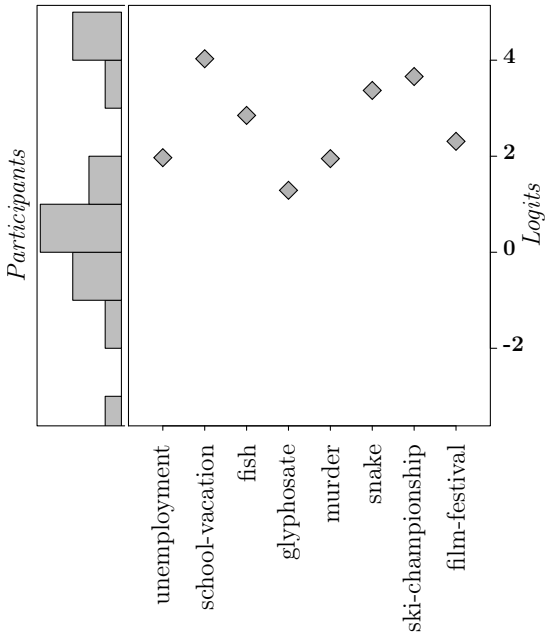


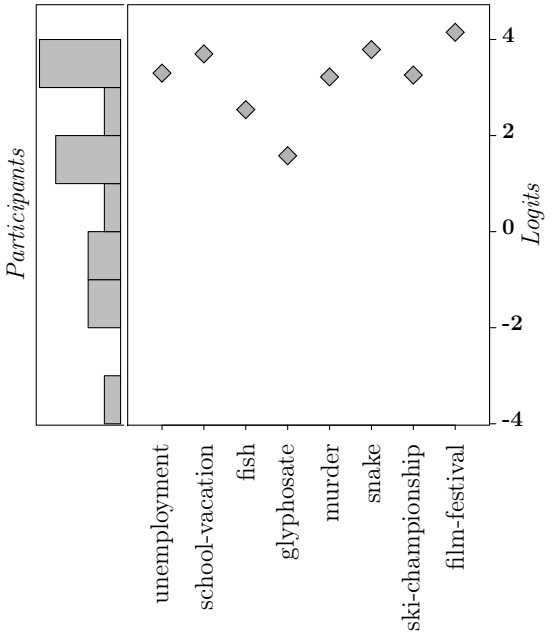
Figure 4.1: Wright map showing participant and item measures for multiple-choice question responses



(a) Easiness of text



(b) Easiness of questions



(c) Enjoyment

Figure 4.2: Wright maps showing participant and text measures for rating responses

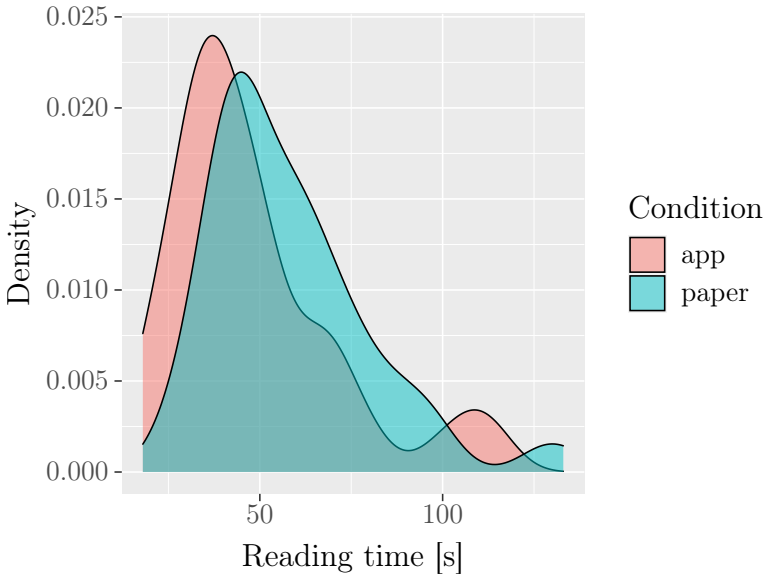


Figure 4.3: Density graph comparing reading times of texts read on the tablet and on paper

4.4.2 Low-level cognitive tasks

Figure 4.4 shows the most relevant scores and reaction times per participant for each of the low-level cognitive tasks.

In the lexical decision task, participants 3, 5 and 10 responded with "WORD" to all stimuli (one of them also in practice task) and did not exhibit any reaction time difference between words and pseudowords. The data of these participants is excluded from this task's data analysis, under the assumption that the task was not correctly understood. All other participants had at least 15 out of 20 responses correct and have consistently slower average reaction times for pseudowords than for words. There is a highly significant correlation between the reaction times in the lexical decision task and the reading time in the comprehension task. The full correlation matrix including all relevant measurements from the different tasks can be seen in Appendix A.

In the short-term memory task, five participants repeated the practice task once, two repeated it twice. As the task allowed only a single attempt and stopped immediately after the first mistake, I also include the practice attempts in the analysis for this task and use the highest score reached among all attempts as the final score.

4.5 Discussion

Despite most comprehension questions being too easy for many participants, the Rasch model still shows a fairly reliable separation between items, and the zero-separation measured between the two conditions can be interpreted as evidence against a significant effect of testing modality. The difficulty and enjoyment ratings, however, separate the

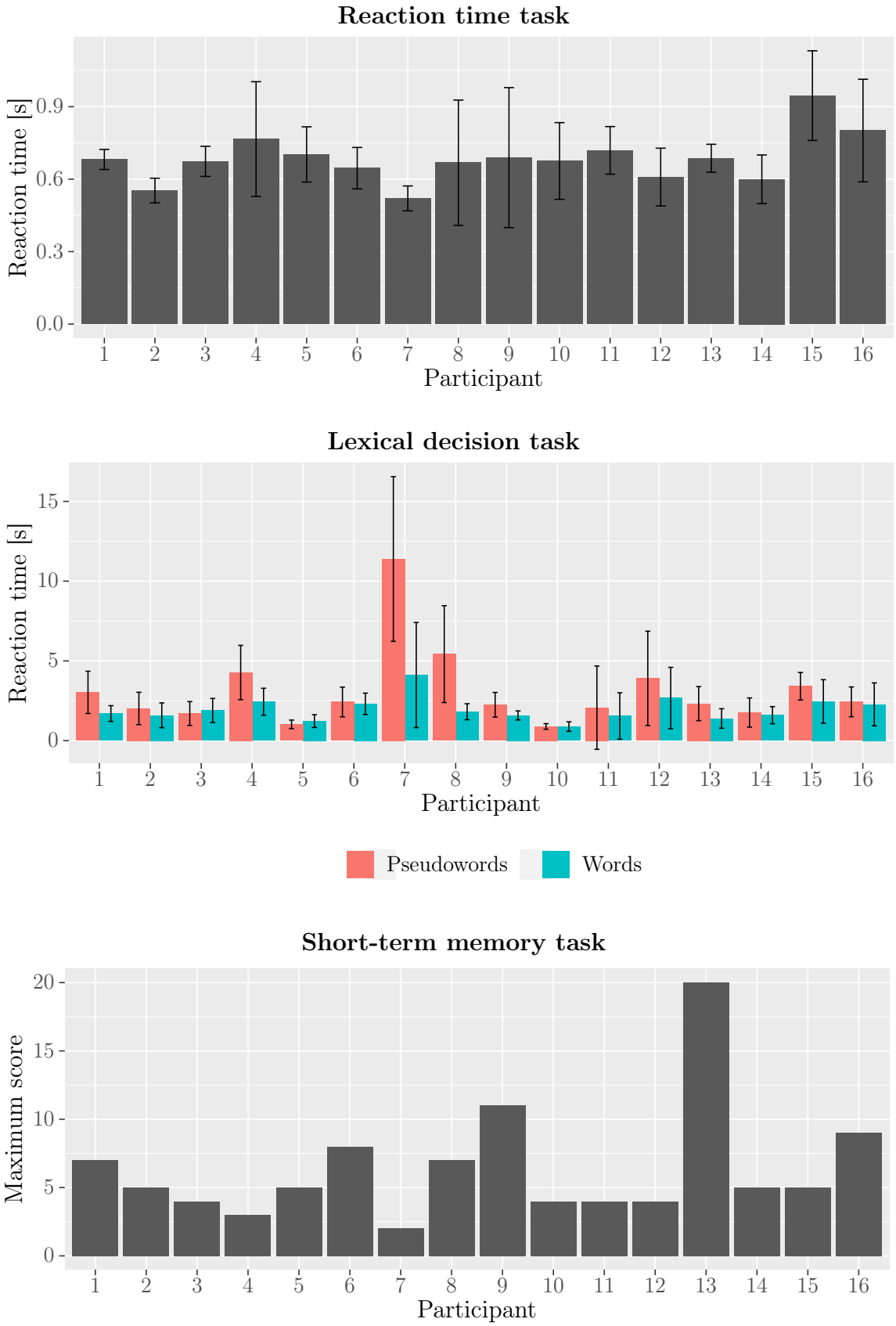


Figure 4.4: Results from low-level cognitive tasks (error bars are standard deviations)

texts to a much lesser degree and are therefore less reliable. Also, it is important to note that some of the highest performing participants rated the texts as very difficult and vice-versa. This illustrates again that subjective perception of difficulty and comprehension are very different things, and that comparing directly between ratings or item accuracies across several participants without correcting for leniency and proficiency is dangerous.

What seems particularly interesting though is the difference in measured reading time. The pure reading time could not be measured for texts read on paper because the ratings were on the same page as the text and there was no way for the experimenter to see when exactly the participant had finished reading, so all measurements also include the time taken for the text difficulty rating. While it is possible that those ratings took longer on paper, or that measurements were less accurate when done manually by the experimenter, these two limitations alone cannot explain the entire difference of almost 10 seconds. One might hypothesize that participants felt more confident reading the texts on the tablet or that there is a lower inhibition threshold when interacting with visual elements on a touchscreen rather than physical objects, but to investigate this phenomenon further, more precise behavioral observations would be needed. However, this apparently did not affect performance in the comprehension questions. Being able to measure time and behavior accurately and precisely is a clear advantage of computerized testing and can be exploited much further than discussed here.

One of the main reasons why one might worry about using a touchscreen device for measuring comprehension is that participants who are more familiar with technology and how UIs work may have an unfair advantage. Observations during the experiment, the high number of restarted practice trials and the high between-subject variability suggest that the short-term memory task was the most difficult in terms of usability and that scores may be affected by varying levels of confidence in handling a touchscreen interface and interpreting visual feedback. For example, the exceptionally well-performing participant achieving a score of 20 in the short-term memory task mentioned that they play video games and that they are familiar with similar types of activities. In contrast, between-subject variability in the reaction time task is quite low and values are within an expectable range, suggesting that there was no major effect of usability in this case. In the reaction time task, there is also no correlation with comprehension task measurements, while the short-term memory task may share some common effects with other tasks (whether it be memory capacity, touchscreen usage skills, or somethinig else).

All this amounts to the conclusion that skills in using touchscreen devices should always be tested, either with more detailed questionnaires about participant's usage of these devices, or more objectively with a separate task. However, in this study, there is no evidence that measuring comprehension using multiple-choice questions is affected by the modality of text and question presentation.

4.6 Limitations

The small sample size is the most obvious limitation of this study and limits its statistical power considerably. Moreover, the sample may be unbalanced due to convenience

sampling. Most participants were very familiar with touchscreen devices and did not have many specific needs in terms of accessibility, which limits the transferability to a more general population.

This also limits conclusions about the usability and accessibility of the application. It has become apparent that at least some of the task implementations are not yet fully optimized for this target group, not to mention that the application is completely untested with other target groups. Systematic usability testing, also including users with more severe intellectual and motor disabilities, is still necessary.

Finally, this experiment was conducted under laboratory conditions in a highly controlled environment. In some cases, particularly in the low-level cognitive tasks, intervention and further instructions from the experimenter were still needed. Taking experiments like this outside the laboratory to simplify data collection even more (as originally intended) would first require further testing and improvement.

5 Conclusion and outlook

I have presented a touchscreen application which implements several well-established methods of comprehension testing in a way that it could simplify and accelerate data collection, especially with groups of participants with reading difficulties. The validation experiment has shown that people with intellectual disabilities can successfully use the application and that results are equivalent to using paper-and-pencil methods. However, further development and testing needs to be done to improve accessibility to a point where the application could be used completely independently and outside of laboratory conditions by diverse user groups.

In future research, the use of this or similar applications could offer considerable benefits, such as reaching more participants, studying comprehension in more natural reading environments, or simplifying longitudinal studies for measuring improvements in reading skills over time. In addition, administration costs can be reduced by having a single mobile application for a variety of experimental tasks which can be updated and communicated with remotely. If accessibility is sufficient, it can also provide a means to more reliable and fine-grained quantitative evaluation of text simplification directly with target groups of easy-to-read language.

Despite this overall successful first test of the application, there are still many open questions: Which types of tasks should be used for different target groups? Which tasks are least affected by the participants' skills at using a touchscreen device? What effect do different reading devices have on reading time? What influence would non-laboratory conditions have on the measurements? Answering these questions would require further experimental research with larger sample sizes.

References

O. Alonzo, J. Trussell, B. Dingman, and M. Huenerfauth. Comparison of methods for evaluating complexity of simplified texts among deaf and hard-of-hearing adults at different literacy levels. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2021.

R. C. Anderson and A. Davison. Conceptual and empirical bases of readability formulas. *Center for the Study of Reading Technical Report; no. 392*, 1986.

A. Bailin and A. Grafstein. The linguistic assumptions underlying readability formulae: A critique. *Language & Communication*, 21(3):285–301, 2001.

A. Bailin and A. Grafstein. *Readability: Text and context*. Springer, 2016.

R. G. Benjamin. Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, 24(1):63–88, 2012.

C.-H. Björnsson. *Läsbarhet [Readability]*. Liber, 1968.

A. K. Boukhvalova, E. Kowalczyk, T. Harris, P. Kosa, A. Wichman, M. A. Sandford, A. Memon, and B. Bielekova. Identifying and quantifying neurological disability via smartphone. *Frontiers in Neurology*, 9:740, 2018.

E. Charzyńska and Ł. J. Dębowski. Empirical verification of the polish formula of text difficulty. *Cognitive Studies*, 15, 2015.

K. Collins-Thompson and J. P. Callan. A language modeling approach to predicting reading difficulty. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 193–200, 2004.

S. A. Crossley, H. S. Yang, and D. S. McNamara. What's so simple about simplified texts? a computational and psycholinguistic investigation of text comprehension and text processing. *Reading in a Foreign Language*, 26(1):92–113, 2014.

E. Dale and J. Chall. A formula for predicting readability. *Educational Research Bulletin*, pages 37–54, 1948.

S. Deilen. Visual segmentation of compounds in Easy Language: Eye movement studies on the effects of visual, morphological and semantic factors on the processing of German noun-noun compounds. In S. Hansen-Schirra and C. Maaß, editors, *Intralingual*

Translation Into Easy Language – or How to Reduce Cognitive Processing Costs, pages 241–256. Frank & Timme, 2020.

S. Dufau, J. A. Duñabeitia, C. Moret-Tatay, A. McGonigal, D. Peeters, F.-X. Alario, D. A. Balota, M. Brysbaert, M. Carreiras, L. Ferrand, et al. Smart phone, smart science: how the use of smartphones can revolutionize research in cognitive science. *PloS one*, 6(9):e24974, 2011.

I. Fajardo, V. Ávila, A. Ferrer, G. Tavares, M. Gómez, and A. Hernández. Easy-to-read texts for students with intellectual disability: linguistic factors affecting comprehension. *Journal of Applied Research in Intellectual Disabilities*, 27(3):212–225, 2014.

L. Feng. *Automatic readability assessment*. PhD thesis, City University of New York, 2010.

R. Flesch. *How to write plain English*. Harper and Brothers, 1979.

J. M. Fletcher. Measuring reading comprehension. *Scientific Studies of Reading*, 10(3): 323–330, 2006.

A. Franzen. Antwortskalen in standardisierten Befragungen [Response scales in standardized surveys]. In *Handbuch Methoden der empirischen Sozialforschung*, pages 701–711. Springer, 2014.

G. Fulcher. Text difficulty and accessibility: Reading formulae and expert judgement. *System*, 25(4):497–513, 1997.

S. M. Fulmer, S. K. D'Mello, A. Strain, and A. C. Graesser. Interest-based text preference moderates the effect of text difficulty on engagement and learning. *Contemporary Educational Psychology*, 41:98–110, 2015.

A. V. González-Garduño and A. Søgaard. Using gaze to predict text readability. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 438–443, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5050. URL `https://www.aclweb.org/anthology/W17-5050`.

A. C. Graesser, D. S. McNamara, M. M. Louwerse, and Z. Cai. Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2):193–202, 2004.

R. Gunning. The technique of clear writing. 1952.

S. Gutermuth. *Leichte Sprache für alle? Eine zielgruppenorientierte Rezeptionsstudie zu Leichter und Einfacher Sprache [Easy-to-read language for everyone? A target group oriented reception study of "Leichte Sprache" and "Einfache Sprache"]*, volume 5. Frank & Timme, 2020.

M. Huenerfauth, L. Feng, and N. Elhadad. Comparing evaluation techniques for text readability software for adults with intellectual disabilities. In *Proceedings of the 11th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 3–10, 2009.

S. Jongstra, L. W. Wijsman, R. Cachucho, M. P. Hoevenaar-Blom, S. P. Mooijaart, and E. Richard. Cognitive testing in people at increased risk of dementia using a smartphone app: the iVitality proof-of-principle study. *JMIR mHealth and uHealth*, 5(5):e68, 2017.

S. Kandula and Q. Zeng-Treitler. Creating a gold standard for the readability measurement of health texts. In *AMIA Annual Symposium Proceedings*, volume 2008, page 353. American Medical Informatics Association, 2008.

E. Keuleers and M. Brysbaert. Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, 42(3):627–633, 2010.

W. K. Kirchner. Age differences in short-term retention of rapidly changing information. *Journal of Experimental Psychology*, 55(4):352–358, 1958.

G. R. Klare et al. *Measurement of readability*. Iowa State University Press, 1963.

A. N. Kluger and S. Adler. Person- versus computer-mediated feedback. *Computers in Human Behavior*, 9(1):1–16, 1993.

G. Leroy, J. E. Endicott, D. Kauchak, O. Mouradi, and M. Just. User evaluation of the effects of a text simplification algorithm using term familiarity on perception, understanding, learning, and information retention. *Journal of Medical Internet Research*, 15(7):e144, 2013.

J. M. Linacre. *Many-faceted Rasch measurement*. PhD thesis, The University of Chicago, 1989.

J. M. Linacre. Facets computer program for many-facet rasch measurement, 2020. URL `https://www.winsteps.com`. Version 3.83.4.

H. Maeda. Response option configuration of online administered likert scales. *International Journal of Social Research Methodology*, 18(1):15–26, 2015.

S. Pappert and B. M. Bock. Easy-to-read German put to the test: Do adults with intellectual disability or functional illiteracy benefit from compound segmentation? *Reading and Writing*, pages 1–27, 2019.

R. Perkuhn, C. Belica, M. Kupietz, H. Keibel, and S. Hennig. DeReWo: Korpusbasierte Wortformenliste [DeReWo: Corpus-based word form list]. 2009.

E. Pitler and A. Nenkova. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 186–195, 2008.

E. Redmiles, L. Maszkiewicz, E. Hwang, D. Kuchhal, E. Liu, M. Morales, D. Peskov, S. Rao, R. Stevens, K. Gligorić, S. Kross, M. Mazurek, and H. Daumé III. Comparing and developing tools to measure the readability of domain-specific texts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4831–4842, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1489. URL `https://www.aclweb.org/anthology/D19-1489`.

A. Säuberli, S. Ebling, and M. Volk. Benchmarking data-driven automatic text simplification for German. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 41–48, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-45-0. URL `https://www.aclweb.org/anthology/2020.readi-1.7`.

G. Stoet. Psytoolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology*, 44(1):24–31, 2017.

E. Sulem, O. Abend, and A. Rappoport. Simple and effective text simplification using semantic and neural methods. *arXiv preprint arXiv:1810.05104*, 2018.

C. Timmers, A. Maeghs, M. Vestjens, C. Bonnemayer, H. Hamers, and A. Blokland. Ambulant cognitive assessment using a smartphone. *Applied Neuropsychology: Adult*, 21(2):136–142, 2014.

S. Vajjala and I. Lucic. On understanding the relation between expert annotations of text readability and target reader comprehension. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 349–359, Florence, Italy, Aug. 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4437. URL `https://www.aclweb.org/anthology/W19-4437`.

S. Vajjala, D. Meurers, A. Eitel, and K. Scheiter. Towards grounding computational linguistic approaches to readability: Modeling reader-text interaction for easy and difficult texts. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 38–48, Osaka, Japan, Dec. 2016. The COLING 2016 Organizing Committee. URL `https://www.aclweb.org/anthology/W16-4105`.

N. Valliappan, N. Dai, E. Steinberg, J. He, K. Rogers, V. Ramachandran, P. Xu, M. Shojaeizadeh, L. Guo, K. Kohlhoff, et al. Accelerating eye movement research via accurate and affordable smartphone eye tracking. *Nature Communications*, 11(1):1–12, 2020.

W. Xu, C. Napoles, E. Pavlick, Q. Chen, and C. Callison-Burch. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415, 2016.

Y. Yuan, K.-m. Chang, J. N. Taylor, and J. Mostow. Toward unobtrusive measurement of reading comprehension using low-cost eeg. In *Proceedings of the Fourth International Conference on Learning Analytics and Knowledge*, pages 54–58, 2014.

A Correlation matrix

Task	Measurement	Comprehension questions: Accuracy	Comprehension questions: Reading time	Comprehension questions: Answering time	Reaction time: Reaction time	Lexical decision: Accuracy	Lexical decision: Word reaction time	Lexical decision: Pseudoword reaction time	Short-term memory: Score
Comp. questions	Accuracy		-0.16	-0.19	-0.12	-0.04	-0.32	-0.28	**0.45**
	Reading time	-0.16		**0.70***	0.11	**-0.56***	**0.81***	**0.71***	-0.23
	Answering time	-0.19	**0.70***		-0.21	**-0.54**	**0.64***	**0.62***	-0.40
Reaction time	RT	-0.12	0.11	-0.21		-0.06	-0.15	-0.32	0.12
Lexical decision	Success rate	-0.04	**-0.56***	**-0.54**	-0.06		-0.22	-0.26	**0.50**
	Word RT	-0.32	**0.81***	**0.64***	-0.15	-0.22		**0.86***	**-0.51**
	Pseudoword RT	-0.28	**0.71***	**0.62***	-0.32	-0.26	**0.86***		-0.39
Short-term memory	Score	**0.45**	-0.23	-0.40	0.12	**0.50**	**-0.51**	-0.39	

Correlation coefficients are Pearson's r, **bold print** means $p < 0.1$, * means $p < 0.05$. Sample size is 16, except for the lexical decision task, where the data from three participants has been removed.

Table A.1: Correlation matrix of measurements from the experiment